摘要

2020 年初，新冠病毒席卷全国。在新冠甲类管理时期，为了及时、有效地阻断病毒传播，相关部门会收集个人信息，并对确诊患者、密切接触者等人群的行动轨迹进行溯源，以此尽可能地找出所有风险人群进行治疗、隔离，然而公众规避披露个人防疫信息的事件屡屡发生。公众如果不按要求据实披露相关信息，会给防疫工作带来极大的困扰和阻碍，在社会上造成极其恶劣的影响。除了新冠肺炎，人类还遭受了多起突发公共卫生危机，如 2003 年的非典、2009 年的甲型 H1N1 流感等，这表明突发公共卫生危机的治理应当受到重视。因此，本研究以新冠疫情为例，探索突发公共卫生危机下公众规避披露个人防疫信息行为的影响因素，并针对疫情防控过程中出现的公众规避披露个人防疫信息的问题提出建议，期望减轻甚至消除这些因素对公众披露个人防疫信息的负面影响，推动突发公共卫生危机下相关部门收集公众个人防疫信息工作合理合法地开展，促进治理工作的顺利进行。

本研究首先爬取相关微博评论，然后利用 LDA 主题模型，从评论文本中分析和提取公众规避披露个人防疫信息行为的相关影响因素。再结合计划行为理论和保护动机理论，围绕不同维度的影响因素，构建公众规避披露个人防疫信息行为的影响因素模型。然后依据 LDA 主题提取的结果设计问卷，并通过线上渠道获取数据，利用 SPSS 软件和 AMOS 软件对数据进行信效度分析和结构方程模型验证，最后通过对研究结果进行分析得出结论并提出相应的建议。

本研究的假设检验结果表明：行为态度、威胁风险对公众规避披露个人防疫信息的行为意愿有显著的直接正向影响作用；反应成本对公众规避披露个人防疫信息的行为意愿有显著的直接负向影响作用；虽然主观规范和知觉行为控制对公众规避披露个人防疫信息行为意愿的直接影响不显著，但是二者对行为态度有着显著的直接正向影响作用。同时通过对因子间的影响效应进行分析发现，文中各个维度的因素对公众规避披露个人防疫信息的行为意愿有不同程度的影响，其中行为态度对行为意愿的总影响效应最大，其次依次是主观规范、反应成本和威胁风险，知觉行为控制也对行为意愿有一定的影响但相对较小。

基于以上结论，本文从端正公众规避披露个人防疫信息的行为态度、降低公众所受到规避披露个人防疫信息的主观规范、降低公众对实施规避披露个人防疫信息行为的知觉行为控制、降低公众对披露个人防疫信息可能带来威胁风险的感知、增加公众实施规避披露个人防疫信息行为的成本这5个角度提出了对策和建议，期望为政府应对类似的突发公共卫生危机提供一些参考。

关键词：突发公共卫生危机；新冠；个人防疫信息；信息披露；规避行为

ABSTRACT

At the beginning of 2020, the novel coronavirus swept across the country. During the Class A management period of the COVID-19, in order to timely and effectively block the spread of the virus, the relevant departments would collect personal information and analyzed the action tracks of the confirmed patients, close contacts and so on through data traceability. It could be seen that timely and effective disclosure of personal information by the public was essential to achieve effective prevention and control of the COVID-19. However, during the outbreak of the COVID-19 epidemic, some people escaped disclosure of their personal information. If the public did not disclose relevant information according to the requirements, the safety of people's lives and property would be severely threatened , caused great trouble and obstruction to the epidemic prevention work, and caused extremely bad impact on the society. In addition to the COVID-19, in recent years, human beings have also suffered from a number of sudden public health crises, such as SARS in 2003 and influenza A (H1N1) in 2009. Therefore, this study takes the COVID-19 as an example to explore the influencing factors of the public's behavior of evading the disclosure of personal information under the sudden public health crisis, and puts forward suggestions, in order to deal with similar public health emergencies in the future.

Firstly, this study used the LDA theme model to analyse and extract the relevant influencing factors of the public's evading disclosure of personal information from the relevant microblog comments, and then combined TPB and PMT to build the influencing factor model of the public's evading disclosure of personal information. Then questionnaires were designed on the result of LDA theme model and obtained through online surveys. After that these data were analyzed by SPSS software and AMOS software to obtain descriptive statistics, the reliability and validity of these data, the results of hypothesis testing and the path coefficients between various influencing factors. Finally, personal management suggestions were put forward based on above analyse.

According to the analysis, behavioral intentions are directly, greatly and positively influenced by behavioral attitudes, risks of the disclosure of personal information, while greatly and negatively influenced by the potential cost of avoidance of disclosure. Behavioral attitudes are greatly and directly influenced by subjective norms and perceived behavioral control. Among all the standardized direct effect, subjective norms' direct effect on behavioral attitudes is the most substantial, followed by the direct effect of perceived

behavior control on behavioral attitudes. Compared to other factors, behavioral attitudes have the greatest effect on behavioral intentions, followed by response cost, risks, subjective norms and perceived behavior control one by one. Compared with perceived behavior control, subjective norms are easier to promote people's behavioral intentions than perceived behavioral control. To sum up, the public's behavioral intentions are the most likely to be affected by behavioral attitudes, followed by subjective norms, response costs one by one, while it can not be influenced by perceived behavioral control obviously.

Based on the conclusions of data analyse, personal management suggestions were made from the perspective of correcting the public's attitudes toward disclosure of personal information, reducing the public's subjective norms of evading the disclosure of personal information, reducing the public's perceived behavior control over the implementation of evading the disclosure of personal information, and reducing the public's perception of the threats and risks that the disclosure of personal epidemic prevention information may bring, and increasing the cost of public evading the disclosure of personal information, in order to help the government to respond to the sudden public health crisis.

Key Words: major public health emergencies; the COVID-19; personal epidemic prevention information; information disclosure; behavioral avoidance

目录

第 1 章 绪论

1.1 研究背景及意义

1.1.1 研究背景

2020 年初，新型冠状病毒肺炎疫情席卷了全国，由于新冠病毒传染性很强且变异速度快，各地地方政府陆续采取了重大突发公共卫生事件Ⅰ级响应，整个国家进入高度紧张的抗疫工作中。2020 年 3 月，世卫组织将新冠肺炎称为全球大流行。随后新冠迅速发展为全球性的公共卫生危机[1]。截止到北京时间 2022 年 5 月 6 日 23 时 31 分，全球有超过 5 亿人确诊新冠肺炎，200 多个国家都受到不同程度的影响。新冠肺炎疫情是新中国成立以来影响范围最广、感染人数最多、暴发规模最大的突发公共卫生事件[2]。

在新冠甲类管理时期，为了及时、有效地阻断病毒传播，相关部门会收集个人信息，并对确诊患者、密切接触者等人群的行动轨迹进行溯源，尽可能找到所有的风险人群予以治疗、隔离。公众披露的个人防疫信息包括基本信息（姓名、性别、身份证号等）、地理信息（行动轨迹、家庭地址等）和健康信息（核酸报告、体温等）等多种个人隐私信息[3]。公众可以通过有线上和线下两种方式披露个人防疫信息：线上主要有小程序、健康码等渠道；线下信息披露方式有填写纸质表单等。同时，个人防疫信息披露的对象也大不相同，公众除了要向政府等官方部门提供自己的信息，进出商场和银行等公共场所时也需要披露自己的个人防疫信息。

从新冠疫情防控可以看出，公众及时且如实地披露个人信息对实现有效的防控至关重要。但是自新冠疫情发生以来，部分民众规避披露个人防疫信息的事件屡屡发生。如 2021 年 2 月，一经营小饭桌的男子确诊后瞒报，致 900 余名儿童感染。2022 年 3 月，马某等 3 人从上海某封控小区外出，自驾到青海省西宁市并隐瞒其来自疫情中高风险地区且系密接人员等事实，最终导致 58 人被确诊为新冠肺炎病例，66 人被诊断为新冠肺炎无症状感染者。由这些案例可知，公众如果不按要求据实披露相关信息，会严重威胁到人民群众的生命财产安全，给防疫工作带来极大的困扰和阻碍，在社会上造成极其恶劣的影响。

除了新冠疫情，近年来，人类还经历过多起突发公共卫生危机，如 2003 年的非典和 2009 年的甲型 H1N1 流感等。突发公共卫生事件具有难以控制、爆发性强以及应对周期长的特点，这不仅严重威胁到公众的身心健康，也给政府相关部门的防控工作带来严峻的挑战[4]。各种突发的公共卫生事件不仅对正常的社会秩序和人类的生存

构成严重的威胁，甚至可能引发政治风险、社会风险各种次生危害。非典型肺炎疫情、新冠肺炎疫情等，都说明突发公共卫生事件应当作为现代公共安全治理体系的重点治理对象[5]。因此，研究在疫情防控过程中公众规避披露个人防疫信息行为的影响因素具有重要的现实意义。

1.1.2 研究意义

（1）理论意义

本研究基于计划行为理论和保护动机理论，以新冠疫情为例探索突发公共卫生危机下公众规避披露个人防疫信息行为的影响因素，对于理解疫情防控中公众规避披露个人防疫信息行为的内部动因和外部影响因素具有重要理论意义，对个人信息披露的进一步研究具有一定的启发和指导意义，同时丰富了突发公共卫生危机中公众应对行为的研究。

（2）实践意义

本研究通过研究和分析疫情时期公众规避披露个人防疫信息行为的影响因素，针对新冠疫情防控过程中出现的公众规避披露个人防疫信息的问题提出建议，来减轻甚至消除这些因素对公众披露个人防疫信息的负面影响，有助于改变部分居民不愿意披露个人防疫信息的态度，有利于推动突发公共卫生危机下相关部门收集公众个人防疫信息工作合理合法地开展，促进突发公共卫生危机治理工作的顺利进行。

1.2. 国内外研究现状

1.2.1 国外个人信息的研究现状

通过搜索以及阅读相关文献，笔者发现国外学者在个人信息方面聚焦于研究个人信息披露、个人信息管理、个人信息保护这几个方面的问题。

（一）国外个人信息披露的研究

国外个人信息披露研究的热点集中于社交网络，其次是移动商务平台，其中 Facebook 和 Twitter 作为国外热门的社交平台是研究的焦点。在研究主题方面，相关研究主要从隐私计算、信任、隐私关注等视角展开，也有学者研究了其他因素对个人信息披露的影响。

国外很多学者从隐私计算的角度对用户的信息披露行为进行了研究。Dinev 等人利用隐私计算理论研究发现，互联网用户会在信任、感知收益和隐私担忧之间权衡后决定是否披露个人信息，只有用户对隐私问题的担忧小于用户的信任感和收益感时，披露行为才会发生[6]。Shaw 等研究发现，在移动商务网站能够确保他们隐私安全的前

提下，移动商务网站用户愿意提供他们的个人数据来换取一些好处[7]。Xu 等借用隐私计算理论研究用户在位置服务中的披露行为，发现如果用户认为披露位置数据带来的潜在风险和损失小于披露带来的收益，则他们才会愿意向服务提供商共享位置数据[8]。

信任在信息行为领域也受到了极大的关注，一些外国学者将信任对个人信息披露的影响作为了研究的焦点。Grosso 等发现，客户对零售商及其员工的信任可以在一定程度上减少顾客的隐私担忧，最终对客户披露信息的意愿产生积极影响[9]。Krasnova 等、Liu 等发现，在线社交网络用户对服务提供商的信任显著影响用户的自我披露[10, 11]。Morosan 等人的研究表明，用户对酒店本身的信任会影响用户对酒店线上程序的信任，而用户在线上程序的披露意愿会受到这种信任水平的进一步影响[12]。信任因素除了直接影响个人的信息披露行为外，还可以通过影响其他因素来影响披露行为[13]。例如，Norberg 等通过研究发现，感知风险对披露意愿的负面影响可以通过信任因素而得到缓解，即当消费者信任信息收集者时，消费者感知到的风险就会降低，进而会披露更多的个人信息[14]。学者 Bergström 的研究结果表明，如果用户的信任程度不同，隐私担忧对用户披露行为的影响作用也不同[15]。

许多国外学者研究发现，个人的隐私关注水平对个人信息的披露有重要影响。例如，Bansal 等研究表明，隐私关注是影响个人信息披露意愿的关键因素[16]。Pentina 等人的研究和 Xu 的研究都表明，隐私关注水平高的用户更不愿意披露个人信息[17, 18]。Malhotra 等人通过研究发现隐私关注水平越高的用户会减少其披露行为[19]。Chen 等人和 Wirtz 等人的研究均表明，用户的隐私关注水平越高，其隐私信息的披露意愿越低，并且随着隐私关注水平的提高，用户甚至会采取伪造、修改个人信息等隐私信息披露的应对或规避行为[20, 21]。

除了上述几个视角，也有学者研究了其他因素与人们信息披露之间的关系。例如，Cho 研究发现性别不会显著影响自我披露意愿的程度[22]。Goldfarb 等研究指出，年长的人往往比年轻人更关心隐私[23]。Rui 等研究发现女性会在 SNS 上分享更多照片[24]。Xie 等发现，网络自我表露受到社会文化、网络状况、年龄、性别和个体心理等多种因素的影响[25]。Lee 等研究发现，个人性格特征是影响用户在 SNS 上自我披露的关键因素，更自恋的人会进行更深层次的自我表露[26]。Acquisti 等通过实验得出人们在披露信息时具有“羊群效应”：如果受访者被告知之前的受访者曾披露过敏感信息，他们更愿意披露敏感信息[27]。Melumad 等实地研究结果表明，相比于个人电脑，消费者往往更倾向于通过智能手机披露隐私信息[28]。Baum 等研究指出，人们通常希望隐藏对自己不利的信息[29]。

（二）国外个人信息管理的研究

个人信息管理研究侧重于个人如何收集、组织、存储、检索和利用个人信息[30]。Whittaker 指出，信息保存会受到未来使用该信息的时间以及地点的影响和信息类型的影响，人们必须衡量信息管理的成本和未来使用信息的收益，而且信息利用是管理过

程的核心环节[31]。医疗健康领域主要关注老年人的个人信息管理，如何让老人实现方便有效的在线个人健康信息管理是研究重点[32]。为此，Turner 等建议设计一个能够提供个人健康信息管理的入口网站[33]。Piras 等人认为，通过跟踪、分析和提取患者个人健康信息的规律，患者可以像医生一样了解自己的病情，甚至可以对他们的病情做出预测[34]。

（三）国外个人信息保护的研究

个人信息的保护一直是国外学者关于个人信息问题的研究重点。最早由美国健康、教育与福利部于 1973 年提出“个人信息保护”的概念和理论，以保障消费者的信息得到公平利用[35]。许多学者研究了有关个人信息保护的法律法规，并提出了相应的策略和建议。例如，King 等人比较了美国和澳大利亚的个人隐私信息的法律，提出了适当的隐私原则来保护消费者隐私[36]。Lim 等认为，有必要分别评估消费者不同类型个人信息的价值，以便制定合理的个人信息保护法规[37]。还有一些学者从企业隐私政策的角度出发研究了个人信息保护。Landau 指出，大多数网站和移动互联网应用程序都要求客户签署数据保护协议，用户别无选择且无实际控制权；此外，公布的隐私政策不完善、不规范，导致其难以有效实施，一旦用户遭遇个人数据泄露，就没有相关规定来保护他们的权利[38]。

1.2.2 国内个人信息的研究现状

通过梳理相关文献，笔者发现国内也有不少学者对个人信息进行了相关研究，但是相较于国外国内该方面的研究起步较晚，主要关注各种情境下个人信息保护方面的问题，近些年来也有不少学者进行了个人信息披露方面的研究。

（一）国内个人信息披露的研究

通过梳理、阅读相关文献，笔者发现国内个人信息披露研究的热点集中于社交网络，其次是移动商务平台，其中微博、微信作为国内热门的社交平台是研究的焦点。在研究主题方面，国内也是主要从隐私关注、信任、隐私计算等角度出发，研究个人信息披露的影响因素。

国内大量研究证实，网络用户的隐私关注与其个人隐私信息披露行为有着密切联系。沈旺等认为，社交媒体用户的隐私关注会直接影响其短期和长期披露意图[39]。胡昌平研究表明，隐私关注对虚拟社区用户信息披露行为有影响[40]。张星研究表明，隐私关注影响用户在线健康信息的披露态度[41]。许多学者如费倩、任卓异等、郭海玲等通过研究发现，隐私关注对个人信息披露意愿具有显著的负向影响[42-44]。顾秋阳等通过实验显示，隐私关注对用户自我披露的负面影响会随着时间的推移增加[45]。李惊雷则通过研究发现，性别不同隐私关注对自我披露的负向影响也不同，对女性群体而言，隐私关注对自我披露具有显著负向影响，而对于男性群体而言并不显著[46]。王萩研究

了用户在 SNS 平台上披露个人信息的主要影响因素，发现隐私态度对用户的最终信息披露意向起到决定作用[47]。高锡荣等发现，隐私关注会使网络用户拒绝提供或者提供虚假的个人信息[48]。

以隐私计算为视角，国内学者主要对个人信息披露的收益和成本因素展开了研究。如李海丹等研究得出，感知风险和感知收益是用户隐私披露意愿的强影响因子[49]。张涛研究得出，移动商务用户的个人隐私信息披露意愿受感知收益的正向影响[50]。而且伴随着研究的进一步深入，学者们进一步扩展和深化了隐私计算中风险和收益的内涵。例如，张星等认为在线健康信息服务用户的感知收益包括个性化服务和情感支持[41]。朱鹏等认为，感知收益分为社交价值和信息与工具价值[51]。兰晓霞等认为，感知收益分为社交收益和功利收益[52]。朱侯等将感知收益分为利己收益（内部关注视角）以及利他收益（外部关注视角），并通过研究证明了这两种收益都对信息披露意愿有正向影响[53]。学者张雯将感知成本划分为客观成本与主观成本[54]。

国内多数学者研究发现信任与个人信息披露意愿呈正相关。如费倩研究发现，消费者信任显著正向影响其信息披露意愿[42]。王雪妍研究发现，用户对移动办公应用的制度信任与其个人信息披露意愿呈显著的正相关[55]。王瑜超在研究医疗社区网站的个人健康信息披露行为时指出，用户对网站的信任和对医生的信任是用户信息披露的前提[56]。但是也有少数学者研究得出信任与个人信息披露行为之间不存在相关关系。如姒琪莹研究表明，用户对 SNS 网站信任度与用户信息披露行为之间并不相关[57]。

自新冠疫情暴发以来，我国需要收集相关公众的个人信息以应对疫情，但是全国多地都出现了公众规避披露个人防疫信息的事例，因此有学者就新冠疫情时期公众的个人信息披露进行了研究。张宇栋等探究了在新冠疫情常态化形势下的社区治理中居民个人信息披露意愿，研究表明居民更愿意在利众和高安全相关度的场景下披露人脸识别等个人信息[58]。李欣松在自分类理论的基础上，探索了公众在信息公开过程与信息收集过程下不同的隐私决策倾向，研究表明人们的信息提供意愿既会受到公民身份下的公共利益和隐私管理担忧的影响，也会受到个人身份下的个人利益和个人隐私担忧的影响[59]。池毛毛等重点研究了疫情初期湖北返乡人员对基层政府的信任对信息不披露意愿的影响，发现疫情下民众对基层政府的信任程度越高越愿意进行个人信息披露[60]。

（二）国内个人信息保护的研究

国内学者十分关注个人信息保护的层面的个人信息问题。如史卫民以个人信息保护法立法为视角分析了个人信息保护存在的风险，并建议要出台一部具有绝对权威性的个人信息保护法[61]。齐爱民指出在中国同样需要维护个人信息，并且应该建立对隐私的法律保护[62]。刘百灵等认为，隐私政策可以让用户信任平台并缓解用户对个人隐私泄露的担心，这是保护隐私的重要途径也是用户法律维权的重要保障[63]。尹异凡认为平台缺乏隐私信息管控和隐私信息保护、用户缺乏个人隐私信息保护意识以及法律

政策保护上的不足是导致个人信息泄露的主要原因[64]。

自新冠疫情发生以来，我国许多学者研究了该背景下个人信息保护的问题。如占南基于隐私保护设计理论研究了重大疫情防控中的个人信息保护问题，认为疫情防控背景下对个人信息的保护涉及多种利益主体，在重视对个人隐私信息的收集和利用的同时，还要兼顾经济发展和个人隐私保护等需求[65]。高志宏从公共利益的层面研究了个人信息保护的问题，指出在我国的个人信息保护法中存在公共利益条款的公共利益内涵和外延模糊、代表机制缺失等问题，这就导致了由于公共利益泛化滥用而侵犯个人信息权益的现象，这些现象在突发公共事件应对中尤为明显[66]。江海洋以比例原则为视角研究了疫情背景下个人信息的保护，指出在衡量信息隐私和公共利益时，应该要明确信息隐私权的社会价值和其公共利益属性，避免发生信息隐私权在与公共利益进行衡量时毫无招架之力的现象[67]。

1.2.3 研究述评

笔者通过查阅相关文献，发现国内外均有许多学者研究个人信息，但是国外在该领域的研究起步较早，主要关注个人信息的管理和保护等方面。而国内学者重点研究了个人信息保护方面的问题。

关于个人信息披露相关的研究，国外主要研究以 facebook 为主的社交平台，随时间推移学者们还增加了对电子商务、在线医疗等平台的研究，研究对象主要为大学生等青年用户。近年来，国内也有越来越多的有关网络用户隐私披露行为的研究，主要是在借鉴国外相关研究视角和方法的基础上对国内网络环境下的隐私信息披露问题进行研究[68]。从研究内容来看，国外主要从隐私关注、信任、隐私计算等角度进行研究，并不断引入了更多的可能影响个人信息披露的其他因素，而国内学者往往以社会学、心理学等学科的研究理论为基础，并将这些理论与现有研究的影响因素相结合，以研究在用户个人信息披露过程中各种影响因素的作用机理[13]。同时国内学者也关注个人信息披露中的隐私保护行为，并结合信息披露行为的影响因素，研究隐私保护行为的成因及对策。

综上，国内外关于网络用户个人信息披露的研究已经取得了一定的成果，但是目前个人信息披露的研究聚焦在网络环境下用户的自愿信息披露行为，缺乏对突发公共卫生危机防控情境下非自愿信息披露行为的研究。而且学者们多是从正面进行研究，缺少对公众个人信息披露的规避行为及成因的研究。自疫情发生以来，全国各地屡次发生公众规避披露个人防疫信息的事件，查明究竟是什么原因导致公众产生这种消极的行为意愿对突发公共卫生事件的有效防控十分重要。虽然有几个学者对疫情下公众个人信息披露意愿进行了研究，但是要么仅从正面研究了公民个人信息披露意愿的影响因素或要么仅重点研究了公众对基层政府的信任对信息不披露意愿的影响。而疫情

时期公众个人防疫信息披露的对象、披露的信息类型以及披露的场景都复杂多样，从单个视角切入恐难以解释清楚公众规避披露个人防疫信息行为的影响机理。因此本文将基于 LDA 主题模型建模结果，结合计划行为理论与保护动机理论，从个人内在动因和外在环境因素全面地研究突发公共卫生危机下公众规避披露个人防疫信息行为的影响因素，以期提出切实有效的建议来减轻甚至消除这些因素对公众披露个人防疫信息的负面影响，提高公众参与突发公共卫生事件防控工作的积极性，提高突发公共卫生事件的防控效率。

1.3. 研究内容与方法

1.3.1 研究内容

本研究首先爬取相关微博评论，然后采用 LDA 主题模型研究方法，从评论文本中分析和提取公众规避披露个人防疫信息的相关影响因素，再结合计划行为理论和保护动机理论，围绕不同维度的影响因素，构建公众规避披露个人防疫信息行为的影响因素模型。然后在 LDA 主题提取结果的基础上设计问卷，并通过线上渠道最终得到 311 份有效问卷，利用 SPSS 软件对数据进行描述性统计分析，用 AMOS 软件对问卷数据进行信效度分析和结构方程模型验证，最后通过对研究结果进行分析得出结论并提出相应的建议。研究总结如下：

第 1 章 绪论

首先，本章分析疫情时期公众个人防疫信息披露的研究背景，阐述该研究的目的，并在其基础上分析该研究的理论意义和实际意义；然后分别介绍国内外相关研究现状；继而分析本文的研究方法以及针对该问题的研究思路；最后对本文可能存在的研究创新点进行阐述。

第 2 章 基础研究理论及技术

本章节对研究的相关概念、方法进行简要概述，对该研究问题涉及的计划行为理论、保护动机理论、LDA 主题模型等基础理论和方法进行简介，同时明确各理论之间的关系以及各理论、方法对于该研究问题的基础性作用。

第 3 章 基于 LDA 模型的影响因素提取

本章主要依据相关关键词爬取微博数据，然后通过人工筛选、构建停用词词典等方式对已获取的微博评论数据进行预处理，接着确定各类参数，并将所有参数代入并运行代码进行 LDA 主题模型训练，最后基于 LDA 主题模型对微博评论文本的主题进行标识和归纳。

第 4 章 研究假设与构建模型

本章以第 3 章主题提取结果为基础，结合计划行为理论和保护动机理论，提出本文的 7 个研究假设，并构建公众规避披露个人防疫信息行为的影响因素概念模型。

第 5 章 数据收集与分析

本章以第 3 章主题提取结果为基础设计自变量部分的初始问卷，同时借鉴成熟量表设计因变量的初始问卷；然后进行预调研，依据预调研结果对问卷进行修正；继而正式发放问卷回收数据进行描述性统计分析、信效度分析以及结构方程模型验证，最后对实证分析结果进行总结，从而分析突发公共卫生危机下公众规避披露个人防疫信息行为的影响因素并总结规律。

第 6 章 研究建议与展望

本章首先结合前面章节的研究，总结出影响突发公共卫生危机下公众规避披露个人防疫信息行为的因素，提出相关的建议；然后对整个研究进行总结；最后指出研究过程中存在的不足与缺点，进而对未来的工作进行展望。

1.3.2 研究方法和技术

（1）文献分析法：文献分析法是研究的基础，本文将通过“中国知网”和“Web of Science”文献数据库检索相关文献，并对收集到的相关研究文献进行梳理和归纳，分析国内外研究现状，为论文的继续研究提供理论依据。

（2）LDA 主题模型：本研究将用 Gibbs 抽样算法选取参数值，对 LDA 算法模型进行了构建，并利用困惑度指数寻找最佳主题数，探索各主题下影响公众规避披露个人防疫信息行为的关键因素。

（3）问卷调查法：本研究将利用问卷调查法收集样本数据并进行处理分析，通过微信、QQ 等线上渠道发放调查问卷，以获取最真实可靠的数据从而解释变量间的相互关系，并利用 SPSS 软件对数据进行信效度分析和描述性统计分析。

（4）结构方程模型：本文根据假设结果构建结构方程模型，利用统计分析软件 AMOS 对正式收集到的问卷数据进行测量模型分析和结构模型分析，并根据数据分析结果厘清公众规避披露个人防疫信息的影响因素。

1.4 技术路线图

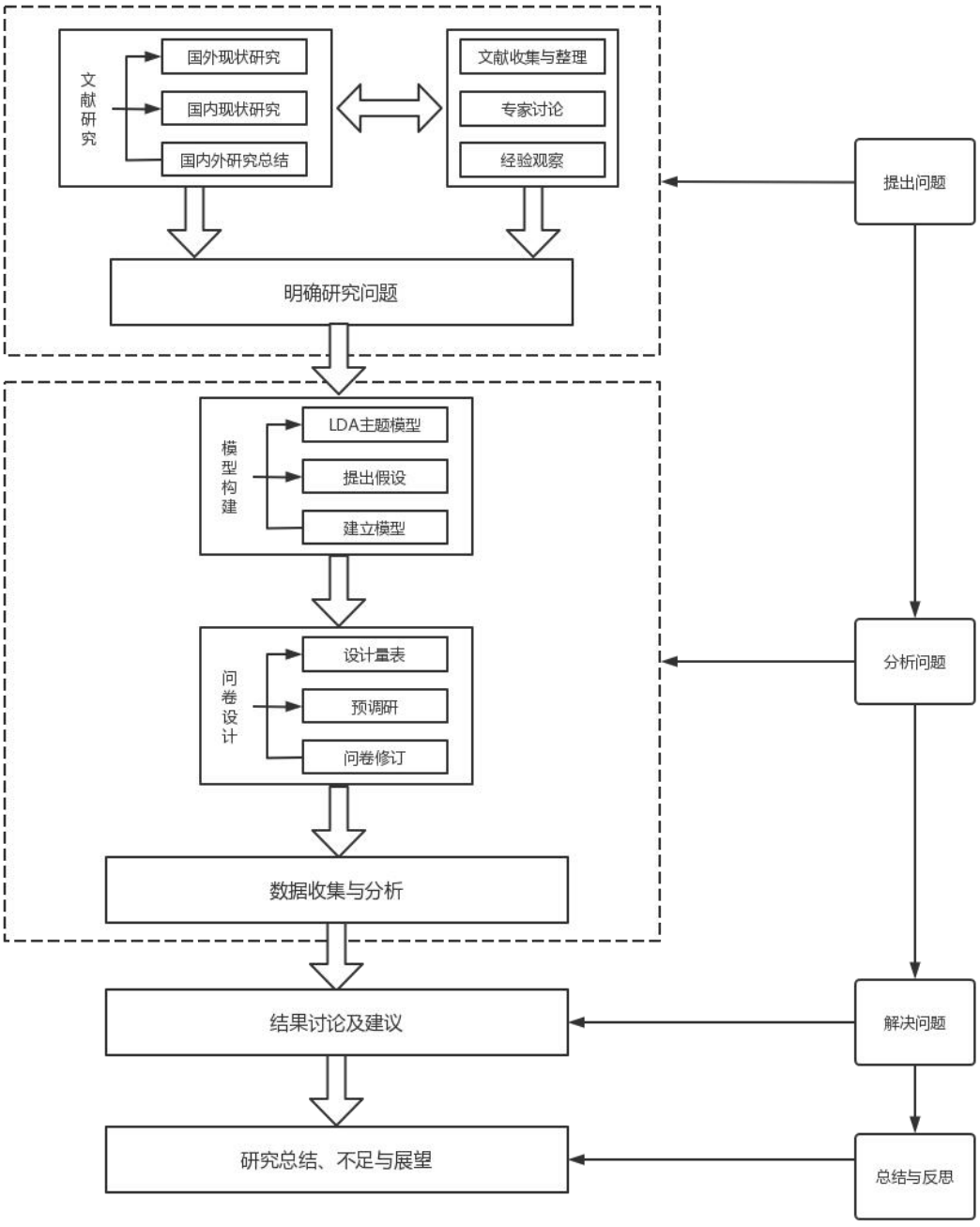

图 1.1 技术路线图

1.5 本文创新点

（1）在研究角度上寻求创新。通过文献梳理发现，目前有学者以公民对基层政府的信任为角度进行了疫情时期公众不披露个人防疫信息意愿的影响因素研究，但是疫情时期公众个人防疫信息披露的对象、披露的信息类型以及披露的场景都复杂多样，因此公众规避披露个人防疫信息的行为可能会受到多方面因素的影响，从单个视角切入恐难以解释清楚公众规避披露个人防疫信息行为的影响机理。因此本文基于计划行为理论和保护动机理论，既考虑了个人内在动因也考虑了外在环境因素对公众规避披露个人防疫信息行为的影响，相较于目前主要从用户感知基层政府作用的角度出发研究疫情时期公众个人防疫信息的不披露意愿，本文研究的角度更全面。

（2）在研究方法上寻求创新。通过文献梳理发现，对个人信息披露意愿的研究大多采用访谈并结合已有量表的方式设计问卷，但是访谈结果的准确性和可靠性可能受到研究者素质等主观因素的影响。同时对公众规避披露个人防疫信息行为的研究属于对人负面行为的研究，但由于访谈不具有匿名性，因此此类涉及访谈对象敏感以及隐私的问题不宜采用访谈的形式。而微博是一个被广泛使用的大型网络社交平台，很多人都会在微博上更为自由地表达自身的看法和意见，因此通过微博可以收集更加直接客观的文本。所以，本文以我国的新冠疫情为研究背景，使用爬虫的方法采集微博数据，利用 LDA 主题建模方法对微博评论数据开展主题识别研究，然后基于 LDA 主题模型的研究结果设计问卷量表，最后使用结构方程的方法对调研数据进行处理分析。因此，本文为个人信息披露的研究提供了一种新思路。

第 2 章 理论基础与关键技术

2.1 理论基础

2.1.1 计划行为理论

Ajzen 和 Fishbein 在 1975 年提出了理性行为理论（TRA），指出行为意愿可以直接决定行为发生与否，并且行为意愿的强弱受到行为态度和主观规范的影响[69]。后来 Ajzen 考虑到人类不可能用自己的意志完全控制自己的行为，因此在理性行为理论的框架下引入了知觉行为控制这一要素，提出了著名的计划行为理论*（Theory of Planned Behavior，TPB）*。根据计划行为理论，个体的行为意愿直接影响个体的行为，而行为意愿则主要受到行为态度、主观规范及知觉行为控制这三个层面的影响[70]。其中，TPB 中的态度是指个体对某种行为所持有的正面或负面的感觉；主观规范是指个体在考虑是否采取某种行为时，所感受到的外部环境所造成的社会压力；知觉行为控制是指一个人对其执行行为时所感知的控制能力[70]。

许多学者基于计划行为理论研究了人各种行为意向的影响因素。刘健等基于计划行为理论对人们的高速铁路乘坐意向进行了研究，发现行为态度、主观规范和知觉行为控制都对公众的高铁乘坐意向有直接正向的影响[71]。马壮林等以计划行为理论和技术接受模型为基础研究了在限行政策下城市居民低碳出行意向的影响因素及其相互作用机理，研究结论可为交通管理部门提高限行政策效益、引导居民低碳出行提供理论支撑[72]。王建华等基于拓展的计划行为理论模型研究发现消费者对安全认证农产品的购买意向受到消费者对农产品价值的感知、行为态度、知觉行为控制以及主观规范的正向影响[73]。王月辉等基于计划行为理论和技术接受模型，研究得出北京居民对新能源汽车购买意向受到购买态度、知觉行为控制以及主观规范的关键影响作用[74]。

2.1.2 保护动机理论

保护动机理论*（Protection Motivation Theory，PMT）*于 1975 年由 Rogers 提出[75]，并于 1983 年被 Maddux 和 Rogers 进一步完善[76]，形成了一个完整的理论。该理论认为，当个体面对内部环境以及外部环境中的某些危险因素时，会以一种有利的或者不利的方式做出反应[77, 78]。该理论表明，威胁评估和应对评估是应对威胁的两个主要应对过程[79]。一个行为的发生不仅需要进行威胁评估来评价威胁的发生概率和严重程度，还需要进行应对评估来评价对威胁的处理能力[80]。其中，威胁评估过程包含两个要素：

感知威胁的可能性、感知威胁的严重性；应对评估过程包含三个要素：反应成本、反应效能和自我效能[81]。

保护动机理论最初来自健康和社会心理学领域，用于解释人们对于自我保护行为的社会认知[82]，后来被越来越多的学者应用于其他行为领域。董云龙基于保护动机理论，探讨了用户避免密码重复使用意愿的影响因素，研究发现，感知威胁的可能性、反应效能、信息安全意识对他们避免密码重复使用的意愿具有显著影响；而感知威胁的严重性、自我效能、感知收益、反应成本对该行为意向并无显著影响[82]。潘越基于保护动机理论和计划行为理论对城市自行车骑行者不安全骑行行为进行了研究，发现城市居民的不安全骑行行为的行为意向较大程度地受到行为态度、威胁风险、知觉行为控制和主观规范的影响[83]。陈宇琨基于保护动机理论，用威胁评估和应对评估的概念和产生保护动机的过程来解释疫情对个体天人观念的影响，继而进一步解释对消费者可持续购买意愿产生的影响[84]。曾丹以保护动机理论为理论支撑，通过对国内一知名微博辟谣平台的文本展开研究，进而对社交媒体辟谣传播的具体效果进行探讨[85]。

2.2 关键技术

2.2.1 LDA 主题模型

主题模型属于一种非监督的机器学习，是使用算法来对语料库进行聚类从而生成潜在主题的统计模型。而 LDA 模型由 Rashid 等人于 2003 年提出的[86]，是目前运用较为广泛的主题模型之一。LDA 包括词语、主题和文档 3 个层次，其中文档是主题的概率分布，而主题又是词的概率分布。对于每篇文档，LDA 都有一个固定的生成过程：首先对于每篇文档，在文档的主题分布中随机提取一个主题；然后在与所选主题对应的词语分布中随机提取一个词语；重复上述过程，直到最后遍历文档中的每个词。该方法可以将语料库中每个文档的主题以概率分布的形式呈现，可以有效地聚合文本的主题特征，并且对文本的长度没有严格的限制[87]。它通常被用于文本主题的识别、文本聚类和文本特征降维的研究。由于 LDA 主题模型减少了大量文本分类过程中的主观偏见[88]，使研究人员可以找到潜在的主题，而不是把预先建立好的分类强加在数据上，从而改进了学者们在文本数据中思考和解释主题的方式[87]。

LDA 主题模型可以帮助国内外研究人员有效地从文本中提取有价值和可理解的信息，并被广泛应用于在线评论、自然语言处理和文献挖掘等各个领域。在在线评论领域，余佳琪等创建了基于 LDA 的评论主题情感协同挖掘模型，以期能够及时了解慢性病患者在病情不同时期的情感与关注主题[89]。史昀嘉通过实践表明，用微博集预先训练的 LDA 模型，可以相对较好的挖掘微博用户的兴趣偏好[90]。马玉基于 LDA 模

型从在线评论中挖掘潜在的产品创新机会，以期为企业创新提供参考[91]。张梦璐对喜马拉雅 FM 用户评论进行了基于 LDA 主题模型的文本挖掘，并探究了影响用户继续使用知识付费平台意愿的因素[92]。唐子豪通过对评论文本进行了基于改进的 LDA 模型的主题挖掘，来识别垃圾评论，进而为排除垃圾评论对用户的干扰、提升用户消费体验提供了建议[93]。

2.2.2 结构方程模型

结构方程模型（SEM）又被称为协方差结构模型[94]，是一种用于测量不可直接观测变量与观测变量之间关系的统计方法[95]，它基于变量的协方差矩阵来分析变量之间的关系。结构方程模型由测量模型与结构模型组成。测量模型用于揭示潜变量与观测变量之间的关系，其公式如式 2.1 和 2.2 所示：

$$X=\Lambda_x\boldsymbol{\xi}+\boldsymbol{\delta} \tag{2.1}$$

$$Y=\Lambda_y\boldsymbol{\eta}+\boldsymbol{\varepsilon} \tag{2.2}$$

公式 2.1 和 2.2 中，X、Y 分别表示外因观测变量以及内因观测变量；ξ、η则分别表示外因潜变量和内因潜变量；Λ_x、Λ_y 分别表示外因潜变量与其观测变量间的系数矩阵和内因潜变量与其观测变量间的系数矩阵；δ、ε分别表示外因观测变量和内因观测变量的测量误差。

结构模型主要是用于检验潜变量之间的关系，通过路径系数来体现。结构模型的公式如式 2.3 所示：

$$\boldsymbol{\eta}=\mathbf{B}_{\boldsymbol{\eta}}+\boldsymbol{\Gamma\xi}+\boldsymbol{\varsigma} \tag{2.3}$$

公式 2.3 中，B 为内因潜变量的结构系数矩阵；Γ为外因潜变量与内因潜变量的结构系数矩阵；ς表示外因潜变量与内因潜变量间的测量误差。

结构方程模型是一种线性统计建模工具，由于其可对不可直接观测的变量进行复杂因果关系的测量，有效弥补了传统统计方法的不足。在经过长期发展后，它被广泛运用于经济学、行为科学等领域。许多学者利用结构方程开展了对个人信息披露方面的研究。陈思琪分别从社交商务平台、消费者以及消费者与社交商务平台互动这三个维度作为切入点，构建了消费者隐私披露行为影响因素的结构方程模型[96]。李锦辉等利用结构方程模型的方法验证微信用户隐私披露的行为模型，其研究可帮助人们更好地理解中国情境下的隐私犬儒主义和隐私悖论问题，并为政府及其有关部门不断明确政策法规提供方向和指引[97]。李琪等基于隐私计算理论和社会资本理论，构建了移动社交平台隐私披露意愿的影响因素模型，利用结构方程模型进行验证分析[98]。余东辉研究了在我国电子商务环境下影响消费者信息隐私披露行为的因素，最后通过结构方

程建模等方法对收集的 449 份有效问卷进行了分析，并从电子商务企业的角度提出了管理建议[99]。

2.3 理论与技术在本文中的应用

保护行为理论虽然最初用于研究人的各种健康行为，但越来越多的学者通过研究证实该理论同样也可以很好地解释人们在其他领域的行为。在信息安全行为的大背景下，保护动机理论应用较为广泛[82]。因此本研究将公众采取的规避披露个人防疫信息行为视为一种信息保护行为。计划行为理论是目前关于个体行为生成的最重要的理论之一[100]。计划行为理论认为，大多数人都是理性的，影响人们实施一个行为的主要因素是他们相应的行为意愿以及对个人内在因素和外部环境因素的控制能力[101, 102]。而疫情时期公众个人信息不披露的行为是公众个人进行行为决策的过程，因此计划行为理论对本研究具有很好的说服力。保护动机理论认为人会在评估外界威胁以及自身能力等基础上决定是否实施保护行为，而计划行为理论认为人们是否会实施一个行为受到其对个人内在因素和外在环境因素的控制能力的影响。因此这两种理论既有特别之处，又有相似之处，具有很好的契合性。

计划行为理论中的知觉行为控制来自 Bandura 的自我效能理论，指的是个体感知执行某特定行为容易或困难的程度[103]。信息系统方面的研究通常把知觉行为控制等同于自我效能[104]。Fishbein 和 Ajzen 认为自我效能和知觉行为控制在本质上是一样的[105]。潘越和王莉也认为自我效能和知觉行为控制存在概念上的交叉，因此在研究中将两者视为一个因素[83, 106]。因此，本文将自我效能等同于知觉行为控制，当作保护动机理论和自我效能理论的共同变量。同时，保护动机理论中“感知严重性”为威胁所带来后果的严重程度，而“感知易感性”为个人遭受威胁的概率，为了简化模型，可将其视为一个因素，命名为“威胁风险”。

根据上文分析可知，LDA 主题模型在在线评论主题挖掘的研究方面已经十分成熟。本文使用 LDA 主题模型挖掘爬取到的相关微博评论中的主题，初步探索影响公众规避披露个人防疫信息行为的因素。同时根据上文分析可知，结构方程模型在个人信息披露影响因素的研究方面也已经十分成熟。本文对公众规避披露个人防疫信息行为的影响因素研究中的 6 个潜变量分别设置相应的观测变量，最终形成调查问卷，每个观测变量都以李克特五级量表的形式设置选项，通过结构方程模型量化潜变量之间的影响关系以用于后续的分析。

第 3 章 基于 LDA 模型的影响因素提取

通过对相关文献及理论基础的梳理，对疫情时期公众规避披露个人防疫信息行为的影响机理有了初步了解。由于疫情时期公众规避披露个人防疫信息行为属于一种负面行为，出于对自身的保护等原因，以访谈等匿名性较差的手段难以直接得到公众规避披露个人防疫信息的动机。而随着我国各种网络舆情事件数量和热度不断上升，以微博为代表的网络社交平台已经成为危机事件舆论的主要阵地。微博是“新冠疫情”期间互联网用户活动最为活跃的社交平台[107]。在各种发布公众规避披露个人防疫信息事迹的微博下，许多网友都表达了自己对此类事件的看法以及对导致部分人产生这种行为原因的猜测，这些评论也能折射出网友认为哪些因素可能导致自己实施规避披露个人防疫信息的行为。

因此，本章通过爬取相关微博评论，对文本数据进行预处理后，运用 LDA 主题模型挖掘用户重点关注主题，并通过人工编码的方式从中析取公众规避披露个人防疫信息的影响因素，为后续模型的建立奠定基础。

3.1 数据获取

本文使用网页数据爬取工具“八爪鱼采集器”，运用新浪微博的高级搜索功能，搜索关键词为“隐瞒行程”“瞒报谎报行程”“不配合流调”“不按要求报备”“瞒报活动轨迹”“谎报活动轨迹”“瞒报涉疫行程”“谎报涉疫行程”“隐瞒涉疫行程”“瞒报谎报病情”“隐瞒病情”“瞒报谎报旅居史”“隐瞒旅居史”“瞒报谎报接触史”“隐瞒接触史”“瞒报谎报行踪轨迹”“瞒报涉疫信息”“谎报涉疫信息”“拒不接受社区报备”“不报备”，其中“类型”甄别选项处选择“热门”。爬取 2020 年 1 月 1 日至 2022 年 9 月 10 日的热门微博文本数据，爬取的内容包括博文文本、评论文本、评论数和时间、点赞数和转发数等，最终获得微博评论数据 43200 条并将其存为 xlsx 格式文件，爬取的数据形式如图 3.1 所示。

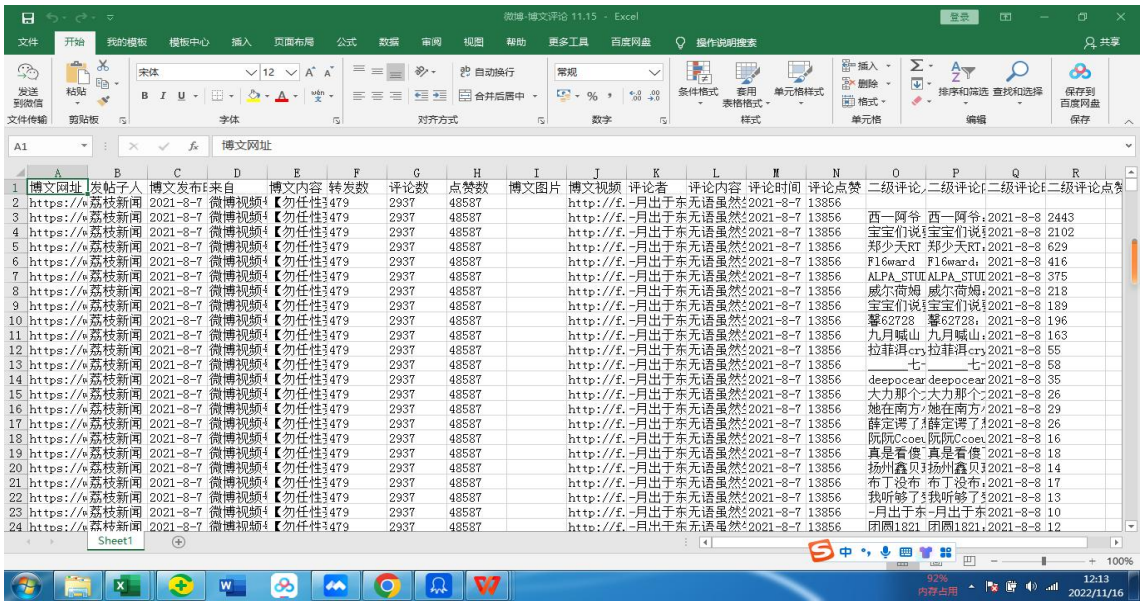

图3.1 爬虫数据一览（部分）

3.2 数据处理

3.2.1 数据预处理

（1）数据清洗与数据筛选

本文微博用户评论文本共 43200 条。因为微博评论文本噪声较大，存在许多重复文本以及纯符号等无意义文本，因此，需要对数据进行清洗。本文运用 Excel 对数据进行清洗，清洗步骤如下：①去除重复文本；②删除评论文本中的@用户名以及带有话题的文本；③删除纯符号等无意义文本；④去除不能体现行为人规避披露个人防疫信息原因的一些评论。首先通过 Excel 的数据去重功能去重后剩余评论文本为 38667 条，再通过数据分列功能将“@用户名：”形式的文本以及带有话题的文本从用户评论中分离出来并且删除，然后通过排序筛选等功能批量删除纯符号和纯表情图等无意义文本，最后通过关键词筛选等方式人工批量去除不能体现行为人规避披露个人防疫信息原因的一些评论。通过以上步骤，最后剩余评论文本 30525 条，部分评论如表 3.1 所示。

表3.1 清洗与筛选后的数据（部分）

评论内容
有一说一，流调出来的个人信息也应该加以法律保护，甚至作为疾控情报定密级。把他人流调隐私发到朋友圈的、泄露给无关人员的，也应当刑拘，该定罪的定罪。这样才能打消配合流调的公民担心自己个人隐私被无端无故泄露的焦虑
你隐瞒什么呢？是有什么见不得人的事吗？

？？是怕小饭桌被查，所以隐瞒？？？？你可以就说聚会聚餐啊……无论怎样，妨害疫情就是更重的罪，害人害己啊

1.无锡新吴区旺庄街道维也纳酒店隔离点，却要收取近万元费用，收费无文件支持，无公示，强制消费。2.这种行为大大增加了老百姓瞒报的风险，与疫情控制的初衷相悖。3.集中隔离点餐食中还出有虫子，发霉的筷子，医护和被隔离人员的食品安全卫生无法保障。???

13 号是不是干了见不得人的事才拒不配合流调

5 天有撒子用？出来照样祸害人，加大处罚

zf 如果向市民积极宣传隔离热线和隔离政策，他们有渠道自行揭发。zf 早说隔离费用不用自费，不就没这些事了么

唉，不工作房贷车贷，什么都还不起

拜托行程卡出人脸识别吧！我真的看到好多抖机灵的事情了！登录别人行程卡就可以逃过一劫！太可怕了！

（2）分词与词性过滤

由于本研究收集的微博数据是以整个文本的形式存储下来的非结构化数据，而在 LDA 主题模型中，语料库中的每个文档都应该是一些无序词语的集合，因此有必要通过中文分词将连续的句子划分为单个词语。而词汇是中文文本的核心，因此分词的结果将直接影响数据处理的有效性。Jieba 是比较常用的中文分词工具，具有易操作、效率高好等优点。Jieba 分词包含精确模式、搜索引擎模式和全模式。在以上三种模式中精确模式最为合适对简短、非结构化的微博评论文本进行分词，因此本研究选择 jieba 的精确模式对文本进行分词。本章研究的目的在于从评论文本中提取网友认为的行为人规避披露个人防疫信息原因的相关主题，为了较好地解读文本的中心思想，尽量减少噪声对重点内容的混淆，本研究对词性进行过滤，仅保留名词和动名词，删除文本分词后的介词、副词等出现频率高但对提取主题没有帮助的词语。

（3）去停用词与词频统计

使用 jieba 对文本分词后，还是存在对文本主题聚类没有帮助却出现频率较高的词以及字，主要包括副词等非名词以及高频却无关的名词，这些词被叫做停用词。而这些词将会影响最终的词频统计等结果，因此需要去除，而是否能有效去除停用词会直接影响到数据处理的结果。因此本研究将目前较为权威的四个停用词典进行整合，包括四川大学机器智能实验室停用词、百度停用词、哈工大停用词以及 GitHub 相关停用词。

由于背景以及需求不同，每个实验需要根据自身情况对停用词词典进行调整，同时还要把对本研究有帮助词语尽量保留下来。因此，借鉴网上的权威的词典初步完成去停用词处理后，本文通过 Python 对数据进行词频统计，来帮助验证预处理结果是否

有效，同时根据词频情况添加停用词以及构建保留词词典，截取的部分词和词频如表 3.2 所示。

表3.2 词频统计（部分）

词	词频	词	词频	词	词频	词	词频
疫情	1917	检测	383	生活	279	数据	223
核酸	1032	防控	380	法律	276	脑子	218
行程	958	政府	367	阳性	272	感觉	218
学校	636	影响	366	传染	272	父母	216
全国	630	人员	364	小区	269	地区	209
自费	605	工作	358	危害	268	轨迹	191
防疫	583	评论	339	社区	266	祸害	191
国家	528	病毒	331	立案	258	商丘	185

3.2.2 数据二次处理

根据词频统计结果显示，分词后的数据还存在很多对文本主题提取无意义的噪声词，如“感觉”“评论”“脑子”等，同时很多可以反映该研究主题的特征词却被分割，这些都将影响主题挖掘结果的精确度。因此本文通过构建停用词、保留词以及同义词词典的方式对数据进行二次处理，具体处理步骤如下：

（1）构建停用词词典

由词频统计的结果可知，对预处理的数据在过滤词性以及去停用词的基础上分词后仍然存在大量对主题提取没有意义的词语，因此，基于词频统计的结果以及研究需要，本研究对已有停用词词典加以补充，新增的部分停用词如表 3.3 所示。

表 3.3 新增停用词（部分）

停用词			
地方	世界	焦作	杠精
评论	问问	看吧	点位
情况	人群	人流	商丘
人员	无语	长长	大门
事情	玩意儿	盲盒	神经
人民	咨询	萝卜	学学
感觉	关注	冲冲	管管

干嘛	个头	荔枝	纯属
垃圾	啥意思	比例	好歹
事儿	大众	开花	做人

（2）构建保留词词典

在本研究的评论文本中含有大量的疫情背景下的专有词汇，如“健康码”“行程码”等，以及一些网络用语如“瓢虫（嫖娼）”等，而 jieba 虽然具有一定的新词识别功能，但由于词汇更新速度快等原因，还是难以准确识别出所有词汇。因此为了确保主题提取的效果，需要根据研究背景构建保留词词典。笔者在结合爬取到的评论文本以及自身经验的基础上，构建了自定义保留词词典，来帮助精确地完成分词任务。本研究添加的主要词汇有“房贷”“隔离”“行程码”“健康码”等，部分保留词词典如表 3.4 所示。

表3.4 保留词词典（部分）

保留词			
行程码 n	手机号 n	生意 n	核酸报告 vn
个人信息 n	手机卡 n	利益 n	传染病防治罪 n
房贷 n	漏洞 n	传销 n	传染病防治法 n
亏心事 n	付款记录 vn	偷渡 n	法律责任 n
防疫政策 vn	消费记录 vn	法律效力 n	批评教育 n
隐私 n	支付记录 vn	层层加码 n	行程码 n
健康码 n	核酸证明 vn	立案调查 vn	法律制裁 n
旅居史 n	生意 n	大数据 n	造谣 n

注：n 表示名词；vn 表示动名词

（3）构建同义词词典

同时由于网友输入粗心等问题，微博评论文本中有许多网络用语以及错别字，如将“核酸”打为“核算”，而这些词语可能会影响主题提取的结果，因此本文借助同义词词典将“核算”转变为“核酸”，同时将“渣浪”转化为“新浪”。

通过上述步骤可以较好地保证本研究中的主题提取结果，但由于文本数据较多等原因，难以通过一次或少数几次的数据处理过程得出较好的提取结果，因此笔者通过多次模型训练不断补充词典，然后进行词频统计，尽可能完善停用词词典、保留词词典和同义词词典，来达到一个较好的训练效果。

3.3 LDA 模型构建

3.3.1 实验参数设定

LDA 主题模型作为一种无监督的机器学习方式，需要人为设定三个参数α、β和 K，其中α和β为狄利克雷分布超参数，默认值分别为 0.37 和 0.02[108]，而 K 为主题数目。对于 LDA 主题模型而言，主题数量密切地关系到文本挖掘的质量。但是目前尚缺乏一个统一的确定 LDA 最佳主题数的办法。困惑度是目前学者们调节主题数目较为常用的一个方法，困惑度可以用来衡量一个概率分布或概率模型预测样本的优劣程度[108]，能够帮助判断概率分布模型训练效果的好坏。而通过主题可视化输出的主题气泡图，可以观察到不同主题的分布情况以及对应的主题频率。因此本研究选用计算困惑度的方法确定主题数目，同时通过观察主题聚类的可视化结果来辅助验证主题数目设置的有效性。困惑度的计算公式如式 3.1 所示。

$$\mathbf{Perplexity(D)} = \mathbf{exp}\ \frac{\sum_{d=1}^{M} \mathbf{logP(W_d)}}{\sum_{d=1}^{M} \mathbf{N_d}} \tag{3.1}$$

公式 3.1 中：D 表示文档中所有词的集合；M 为文档的数量；Wd 为文档 d 中的词；Nd 为每个文档中 d 的词数；P(Wd)为文档中词出现的概率。困惑度为文档 d 所属主题的不确定性，困惑度越小，模型性能越好，困惑度最低或处于拐点时所对应的 K 值为最优主题[108]。本研究分别计算主题数为 1-20 共 20 个模型的困惑度得分以测试最优主题数。实验结果如图 3.2 所示，从图中可以看出，当主题数量在 1 到 7 时困惑度也不断增加，从 7 开始下降，下降到 8 后又开始上升，8 为困惑度最低的拐点，因此本文选取主题数为 8。

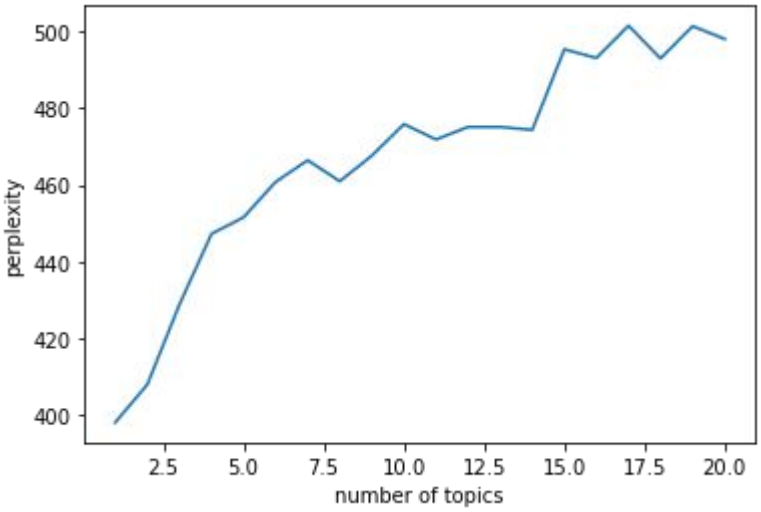

图3.2 困惑度曲线图

3.3.2 主题结果输出与可视化

本文通过一系列步骤完善了停用词、保留词以及同义词词典，并确定了三个需要人为设定的参数的值，然后输入各类词典和参数后运行 Python 代码，输出主题聚类结果及其可视化气泡图分别如图 3.3 和图 3.4 所示。

```
: n_top_words = 15
  tf_feature_names = tf_vectorizer.get_feature_names()
  topic_word = print_top_words(lda, tf_feature_names, n_top_words)
```

```
Topic #0:
防疫 检测 政府 新闻 政策 父母 信息 工作人员 一刀切 志愿者 工作 医护人员 儿子 全员 室友
Topic #1:
隔离 费用 免费 风险 传染 居家 新冠 管控 医疗 买单 记录 自理 疫苗 病情 收费
Topic #2:
医院 惩罚 影响 法律 见不得人的事 后果 孕妇 管理 绿码 检查 抗疫 阴性 开学 黄码 力度
Topic #3:
行程 工作 生活 轨迹 行程码 重罚 患者 法律责任 身份证 监狱 生命 隐私 传染病 影响 一家人
Topic #4:
学校 孩子 学生 家长 家人 大学生 媒体 老师 学籍 教育 大学 上学 事实 同学 处分
Topic #5:
防控 危害 责任 时间 传播 违法 老百姓 单位 弹窗 犯罪 原因 机会 红码 活动 效力
Topic #6:
严惩 流调 回家 侥幸心理 美国 公司 高风险 健康码 医生 旅游 工地 意义 官员 法治 网络
Topic #7:
病例 谎报 大数据 损失 通报 朋友 重判 罚款 措施 经济 群众 物资 浪费 作业 意识
```

图3.3 主题聚类结果

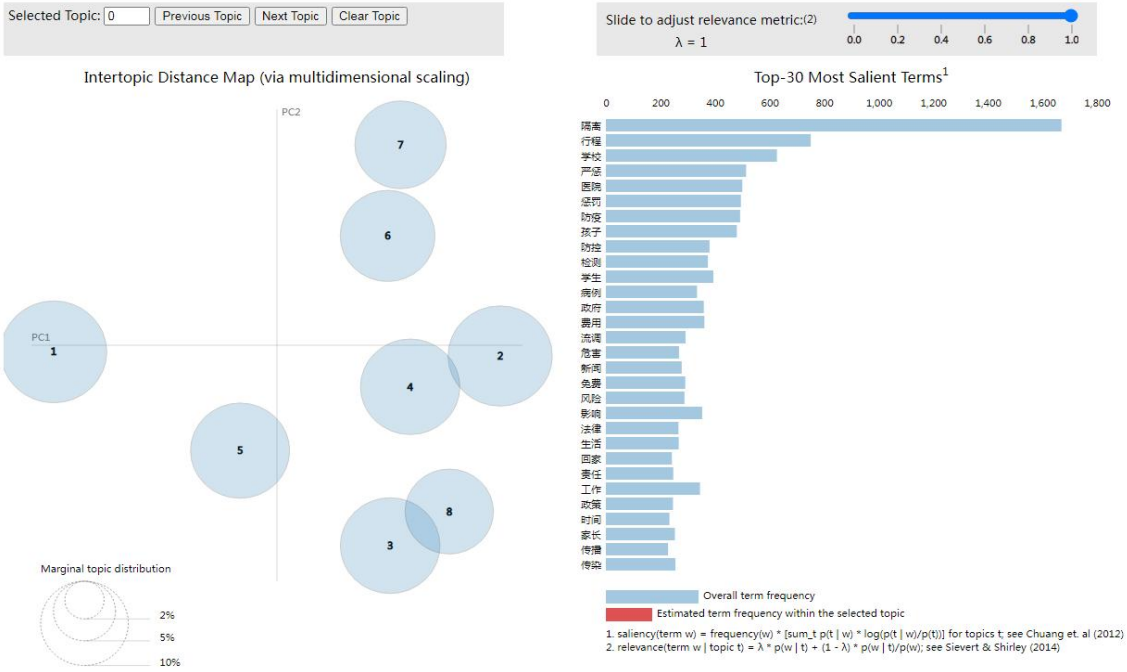

图3.4 主题模型可视化气泡图

如图 3.4 所示，左侧的每个圆圈都代表一个主题，圆圈的相对大小对应了在整个

语料库中主题产生的相对概率。可以看出本次主题聚类虽然存在部分主题与其他主题交叉的情况即主题间存在重复性词语，但整体来看每个主题所代表的圆圈的大小差别不大且即使存在交叉但交叉面积很小，说明聚类结果相对平均且有效，可以为公众规避披露个人防疫信息影响因素的识别、分析提供依据。

3.4 主题标识

为了实现对公众规避披露个人防疫信息行为的影响因素的有效挖掘，本文采用人工筛选无关评论、构建停用词典和保留词典等方式，去除与公众规避披露个人防疫信息行为影响因素不相关的内容，然后确定了一般性参数和主题数目并代入 LDA 主题模型进行训练。为了便于分析，本文选择输出每个主题的前 15 个特征词并进行分析。LDA 主题模型可以帮助提取和理解文档信息，通过结合评论文本对主题聚类结果进行分析，笔者发现一条微博评论往往包含多个对行为人规避披露个人防疫信息原因的描述，一个主题中也包含多方面的影响因素，而且各个主题间虽相互区别但也并非割裂的，这也表明人们实施规避披露个人防疫信息的行为非单一因素作用的结果而受到多方面因素的综合影响。

因此，本文依据这些特征词并结合评论，通过人工编码的方式对主题进行精确地标识，尽可能地提取出每个主题中公众规避披露个人防疫信息行为的影响因素，为后续研究奠定基础。主题分析与识别过程如下文所示：

主题 1 中出现了“防疫”“政府”“政策”“一刀切”等特征词，主要表达了网友认为不信任政府的能力和不认同防疫政策是让人规避披露个人防疫信息的原因，如评论“三年了，还有老百姓隐瞒，是不是政府工作没做到位。让老百姓有后顾之忧？有没有可能政府服务的问题呢？隔离问题呢？不能片面看待问题，旅行也没必要不上报”和“这些现象是屡见不鲜的，背后除了个人的内因，相关防疫措施的落实不统一，一刀切的外因显得更为突出。有时动态真的是动态吗？”。同时该主题还出现了“新闻”“信息”等特征词，一些网友认为由于新闻消息滞后性以及部分民众自身信息闭塞等原因导致了人们不知道自己有感染风险才没有及时上报，如评论“泉州的新闻是 13 号出来的，他是 12 号回来的，应该是以为和疫情错开了，殊不知泉州地方信息可能滞后。”和评论“可是如果不是经常关注这个新闻，是不会知道要报备的……”以及评论“这么多起瞒报的，是因为信息闭塞吗？”。综上，该主题下提取的影响因素有“政策不认同、对政府不信任、新闻滞后、信息闭塞”。

主题 2 中出现了“隔离”“费用”“免费”“买单”“收费”等特征词，不少网友认为一些人因为害怕承担隔离等防疫费用才选择规避披露个人防疫信息，如评论“为什么会隐瞒行程，自费隔离应该是其中一个原因，一天一百对于小老百姓确实不

好受”和“因为隔离费用不低，没钱隔离不起”。该主题中还出现了“风险”“传染”“疫苗”等词，一些网友认为不了解病毒的感染性和危害性是规避披露个人防疫信息的原因之一，如有网友评论“都已经 3 年了，为什么还有人没意识到这个病毒的厉害，一人感染没及时发现可以传染多少人”以及评论“因为打疫苗了，保证死不了”。综上，该主题下提取的影响因素有“费用自付、感知病毒感染性低、感知病毒危害小”。

主题 3 中出现了“医院”“孕妇”等特征词，结合评论发现是网友对“孕妇瞒报”一事的讨论，网友认为该孕妇迫于医生的压力才选择规避披露个人防疫信息，如评论“这是医院不想担责任，怕丢饭碗让孕妇谎报”。该主题中还出现了“惩罚”“法律”“力度”等词，有网友认为法律上惩罚力度小是一些人规避披露个人防疫信息的原因之一，如评论“要是惩罚力度大的话早就没人敢隐瞒了”和评论“看来法律的惩罚力度还是不够啊！判刑 1 年起步，估计敢瞒报的越来越少了”。同时该主题还有“见不得人的事”“影响”等特征词，网友认为一些人由于做了见不得人的事怕曝光了影响自己的名誉，如评论“隐瞒的不是行程，应该做其他见不得人的事情吧？可能行程中有见不得人的事儿，比如约会情人或找小姐 ”综上，该主题下提取的影响因素有“社会支持、刑罚轻、名誉损失”。

主题 4 中“工作”“生活”“一家人”等词体现了网友认为养家糊口的压力以及生活不便会让人铤而走险，如评论“确实，有朋友父母是做摆摊生意的，不出门真的是没有钱进账，儿子还在读大学。遇到疫情隔离了，儿子在学校也隔离没钱吃肉，大小伙子一顿饭就一个素菜。老两口都不知道能不能吃饱饭，这样地找到漏洞一定就会往外跑的，人要生存啊，一家人都要生存啊，对他们来说这已经很难很难了。”和评论“行程为什么要告诉别人呢，不告诉怎么就瞒报违法了呢？那隐私权呢？吃饭，谈恋爱，看电影不算正常生活么？”。另外该主题中“行程”“轨迹”“身份证”“隐私”等词表现出网友对上报后中隐私泄露问题的担心，如评论“以后只会导致更多的人瞒报，因为一旦被发现，所有隐私信息包括身份证门牌号啥啥的全部被泄露，被网爆。谁还敢啊？人家生了病又不是犯了罪，被这样暴露于众”和评论“确实应该配合防疫工作。但是，一旦被定为密切接触者，自己的详细行程都会被曝光”。综上，该主题下提取的影响因素有“生存压力、生活不便、隐私风险”。

主题 5 中出现的“家人”“家长”等特征词主要体现了家人支持对行为人规避个人防疫信息披露的影响，如评论“家长不让他说他能怎么办本来就是家长偷跑的，你指望谁都能有勇气大义灭亲是吧 ”和评论“老太太应该不懂怎么盗码的，应该是有家人教她的”。该主题中的“同学”等特征词则体现了同学的支持对行为人规避个人防疫信息披露行为的影响，如评论“多次隐瞒，同学还帮忙做核酸”。综上，该主题下提取的影响因素有“家人支持、同学支持”。

主题 6 中出现“危害”“违法”“犯罪”等词，体现了网友认为缺乏法律常识是行为人规避披露个人防疫信息的原因之一，如评论“这种人难道不知道瞒报违法吗？”

和评论“应该让大家都知道瞒报的危害和对瞒报的惩治力度”。同时“责任”“单位”等词，体现了网友认为惩罚小不怕影响正常工作也是行为人规避披露个人防疫信息的理由之一，应该把责任落实到单位，如评论“警告？行政拘留？对于事业单位的人来说可能影响工作，当然，他们也不敢瞒报。对于私人公司的或者个体的这种惩罚有何用？就被骂几句而已，不痛不痒，有啥？既然敢瞒报，就不怕几句骂骂！”和评论“我爸就是大学教师，我也在大学工作，我们都签订了承诺书，一旦出现问题，一个系统里的所有上级也是要负责的。为了保全自己的工作确实是两点一线，天天除了单位就是家，更不敢瞒报”。综上，该主题下提取的影响因素有“法律意识缺乏、对工作影响小”。

主题 7 中“严惩”“侥幸心理”等词体现了网友认为行为人有“被发现的可能性”较小的心理，如评论“侥幸心理呗，亏心事不能说，想着万一没事就不会被人发现了”。同时该主题中有“流调”“健康码”等词，网友猜测管理的不到位和健康码漏洞助长了规避披露个人防疫信息事件的发生，如评论“流调工作也有疏忽，那么久都没发现他们儿子”和评论“这其实就是说明现有的防疫措施存在巨大漏洞。南京政府的管理不当，健康码的漏洞等等”。综上，该主题下提取的影响因素有“感知惩罚易感性低、管理疏忽、健康码不完善”。

主题 8 中出现了“大数据”等词体现出网友认为大数据存在不足助长了规避披露个人防疫信息行为的发生，如评论“这说明大数据完全扯淡。去过中高风险地区的检测不到，没去过的瞎弹窗”和“大数据也有准确率问题，用过导航吧，从来没有偏过吗？”。该主题还出现了“朋友”一词，体现了朋友态度对行为人实施规避披露个人防疫信息的行为有影响，如评论“室友怎么想的，好朋友之间虽然感情好还是应该互相正向帮助劝导比较好”。同时该主题中有“损失”“罚款”等特征词，一些网友认为金钱成本低导致有人缺少畏惧心理选择隐瞒行程，如评论“要再加上高额罚款，犯罪成本太低让有些人不当回事”。综上，该主题下提取的影响因素有“大数据不完善、朋友支持、金钱成本低”。

3.5 结果讨论

各个主题标识结果汇总如表 3.5 所示。其中，主题 7 中的“健康码不完善”和主题 8 中的“大数据不完善”都属于防疫技术不完善，因此将其合称为“技术不完善”。同时由于概念相近，把主题 5 中“同学支持”也归到主题 8 中“朋友支持”一类中。综上，本章通过 LDA 主题模型提取出以下 21 个因素：费用自付；生活不便；新闻滞后；信息闭塞；法律意识缺乏；刑罚轻；金钱成本低；对工作影响小；感知病毒危害小；感知病毒感染性低；社会支持；感知惩罚易感性低；家人支持；朋友支持；政策

不认同；政府不信任；技术不完善；管理疏忽；名誉损失；生存压力；隐私风险。

表3.5 LDA主题－特征词分布与主题标识结果

序号	特征词	主题标识
1	防疫 检测 政府 新闻 政策 父母 信息 工作人员 一刀切 志愿者 工作 医护人员 儿子 全员 室友	政策不认同、对政府不信任、新闻滞后、信息闭塞
2	隔离 费用 免费 风险 传染 居家 新冠 管控 医疗 买单 记录 自理 疫苗 病情 收费	费用自付、感知病毒感染性低、感知病毒危害小
3	医院 惩罚 影响 法律 见不得人的事 后果 孕妇 管理 绿码 检查 抗疫 阴性 开学 黄码 力度	社会支持、刑罚轻、名誉损失
4	行程 工作 生活 轨迹 行程码 重罚 患者 法律责任 身份证 监狱 生命 隐私 传染病 影响 一家人	生存压力、生活不便、隐私风险
5	学校 孩子 学生 家长 家人 大学生 媒体 老师 学籍 教育 大学 上学 事实 同学 处分	家人支持、同学支持
6	防控 危害 责任 时间 传播 违法 老百姓 单位 弹窗 犯罪 原因 机会 红码 活动 效力	法律意识缺乏、对工作影响小
7	严惩 流调 回家 侥幸心理 美国 公司 高风险 健康码 医生 旅游 工地 意义 官员 法治 网络	感知惩罚易感性低、管理疏忽、健康码不完善
8	病例 谎报 大数据 损失 通报 朋友 重判 罚款 措施 经济 群众 物资 浪费 作业 意识	大数据不完善、朋友支持、金钱成本低

第 4 章 研究假设与模型构建

经过分析，笔者认为计划行为理论和保护动机理论能够很好地契合 LDA 主题模型的影响因素提取结果。因此，本章以计划行为理论和保护动机理论为研究框架，分析疫情时期公众规避披露个人防疫信息行为的影响因素，提出了 7 个相关的假设，并构建了公众规避披露个人防疫信息行为的影响因素概念模型。

4.1 研究假设

4.1.1 行为态度

根据计划行为理论，行为态度是反映行为个体对于执行某种特定行为的好恶程度的评价[109]。当个体对某种行为的态度越积极，则其采取该行为的意图越高[110]。代志在等人研究表明行为态度对专利信息利用意愿有正向影响[111]。李颖琦等通过对五家大型企业虚拟学习社区知识的行为进行实证研究，发现企业知识共享行为的行为态度会促进其行为意愿[112]。杜旌等人基于计划行为理论研究春节期间返乡行为，发现返乡意愿受到返乡态度的正向影响[113]。杨金铭研究表明态度对居民有机蔬菜溢价支付意愿具有显著的正向影响，而且是影响溢价支付意愿最直接的因素[114]。在本研究中公众规避披露个人防疫信息的行为态度是公众个人对规避披露个人防疫信息行为支持或是不支持的评估，公众对规避披露个人防疫信息行为越支持，则产生规避披露个人防疫信息的意向也就越强烈。因此，本研究提出假设：

H1：行为态度正向影响公众规避披露个人防疫信息行为的行为意愿。

根据以上分析，笔者认为 LDA 主题编码结果中“政策不认同、感知病毒感染性低、感知病毒危害小、对政府不信任”都体现了公众规避披露个人防疫信息的态度，因此在本研究中笔者将用这四个指标作为行为态度的观测变量。

4.1.2 主观规范

根据计划行为理论，主观规范是指个体会受到周围环境及其他人的压力而采取行动，通常周围人对于个体采取行为的期望会影响个体行为[115]。Amalia 运用计划行为理论对共同基金产品的投资意向进行研究，结果表明共同基金的投资意向会受到主观规范的正向影响[116]。魏叶美等对影响教师参与学校治理意愿的因素进行了实证分析，研究表明教师的主观规范对其参与学校治理意愿具有正向的预测作用[117]。王建华等